CSAI 801 Project: COVID-19 Outcome Prediction

Team Names

1. Marim Ashraf Elsayed Mahmoud Amer
2. Mawada Ashraf Elsayed Mahmoud Amer
3. Youmna Alsayed AbdAlatty Mohamed

We have data set preprocessed and it has many features such as location, country, age, and so on. That's data about covid 19 if someone has some symptoms the model will predict the outcome (death/recovered).

In the first, we import all libraries needed in this project and then loaded the data after that we cleaned the data from negative values.

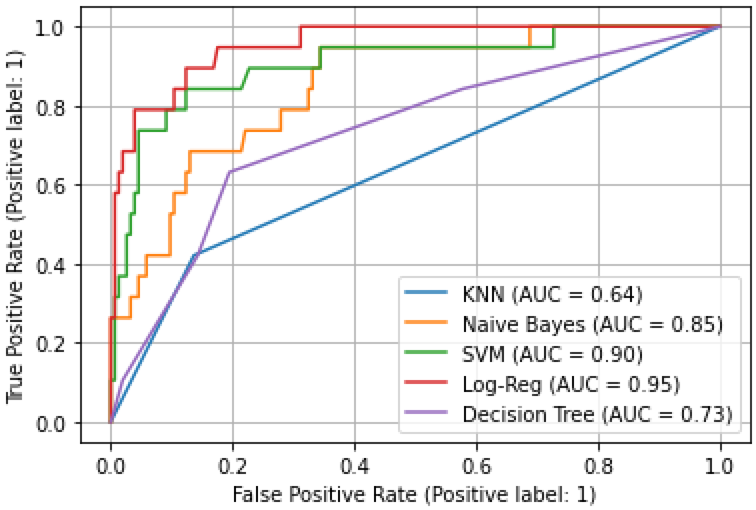
In the second step, we implement one-hot encoding by using the ColumnTransformer function to the data to be easy to manipulate, model, visualize, and create a new binary feature for each possible category and assigns a value of 1 to the feature of each sample that corresponds to its original category. And then we implement the scaling to normalize data to make all data in the same range to preserve important features and this by using the MinMaxScaler function.

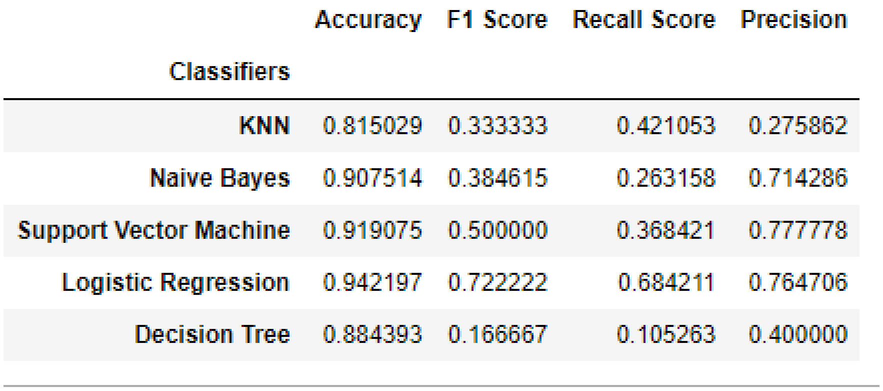
After that, we split data into training and testing data, then splitting training data into training and validation data, we got validation to get the optimal values to hyperparameter to improve accuracy result, then we built different classifier models.

In each model we calculate the accuracy and confusion matrix to know how much the model is good and also use GridSearchCV to be able to get the optimal hyperparameter in each model to get the best accuracy and minimize error.

When we implement optimal hyperparameter in a decision tree, we faced many problems when we write many values to hyperparameter and a lot of features the model takes a long time to calculate. It takes above 2 minutes to run so we decreased the hyperparameter. So we choose only max_depth and max_features to get the optimal hyperparameter.

We implement confusion matrices, f1-score, recall, precision, and accuracy for each model. The best models for these datasets are Logistic and SVM models. Both of them have an accuracy around of 94%.





Classifiers	Accuracy	F1 Score	Recall Score	Precision
KNN	0.815029	0.333333	0.421053	0.275862
Naive Bayes	0.907514	0.384615	0.263158	0.714286
Support Vector Machine	0.919075	0.500000	0.368421	0.777778
Logistic Regression	0.942197	0.722222	0.684211	0.764706
Decision Tree	0.884393	0.166667	0.105263	0.400000

Finally:

From the table:

The **best** model **accuracy** is Logistic Regression

The **best** model **recall** is Logistic Regression

The **worst** model accuracy is KNN

The worst model **recall** is Decision Tree

So the best model is Logistic Regression.